Partially-Aligned Data-to-Text Generation with Distant Supervision

Zihao Fu [1] , Bei Shi [2] , Wai Lam [1] , Lidong Bing [3] , Zhiyuan Liu[4]
[1] The Chinese University of Hong Kong
[2] Tencent AI Lab
[3] Alibaba Group
[4] Tsinghua Univerisity

https://github.com/fuzihaofzh/distant_supervision_nlg

EMNLP 2020

Outline

➡Introduction

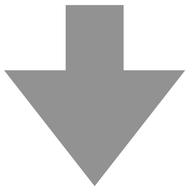
- Dataset
- Model
- Experiments
- Conclusions

Data-to-Text Generation

<Bill Gates, birth_place, Seattle>

Data-to-Text Generation

<Bill Gates, birth_place, Seattle>



Bill Gates was born in Seattle.

Data-to-Text Generation is restricted

- Training generation model needs well-aligned data;
- These data is difficult and expensive to obtain;
- Existing datasets are too small or only contain few particular domains.

Data-to-Text Generation is restricted

- Training generation model needs well-aligned data;
- These data is difficult and expensive to obtain;
- Existing datasets are too small or only contain few particular domains.

Can we use automatically annotated data?

Our Contribution

- We propose a new task: “Partially-Aligned Data-to-Text Generation”;
- We propose a novel framework to make the training dataset automatically;
- We release a partially-aligned dataset WITA;
- We propose a novel Distant Supervision Generation (DSG) framework to utilize the partially-aligned dataset.

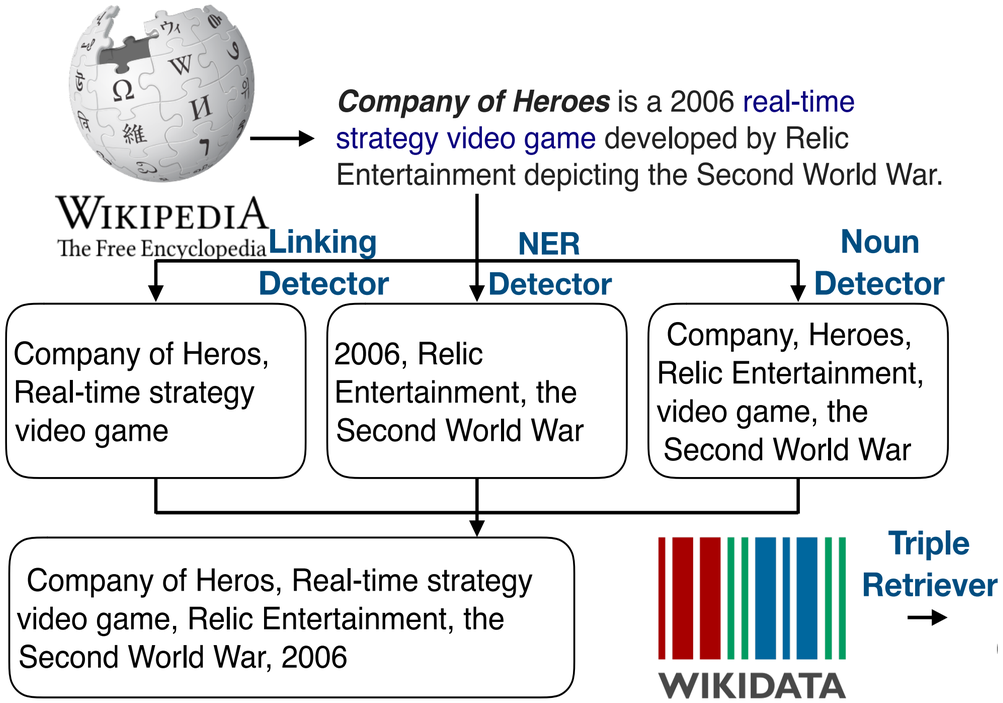
Outline

✓Introduction

➡Dataset

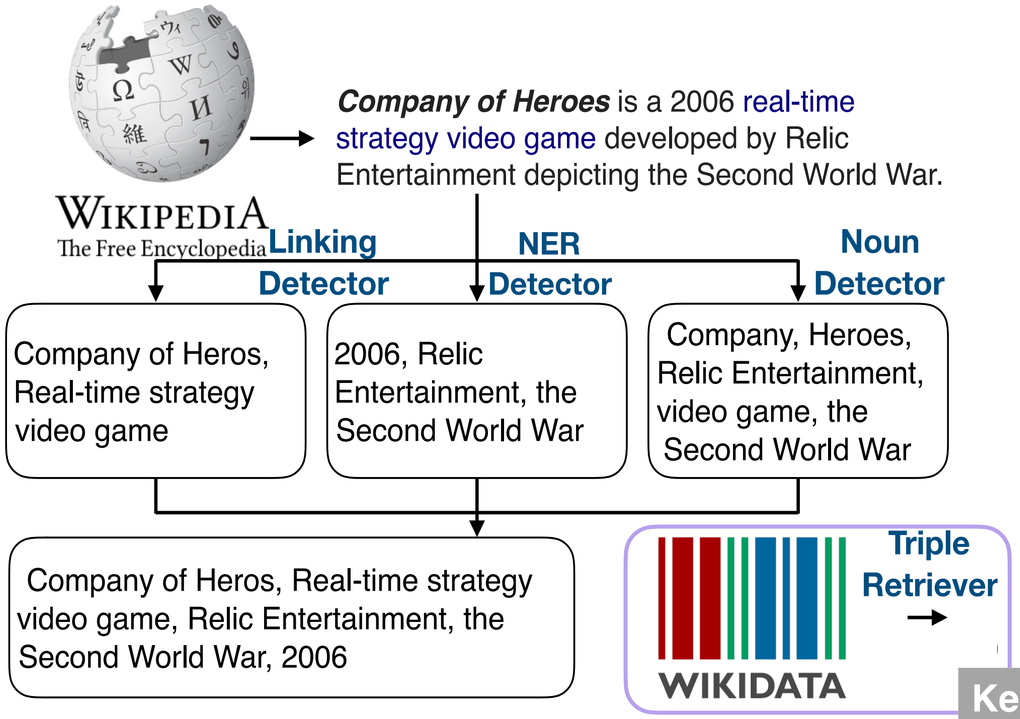
- Model
- Experiments
- Conclusions

Dataset



- Steps:
  - Build Wikidata database.
  - Sample sentences from Wikipedia.
  - Query the Wikidata database to find the corresponding triple for each sentence.

Dataset - Build Wikidata database



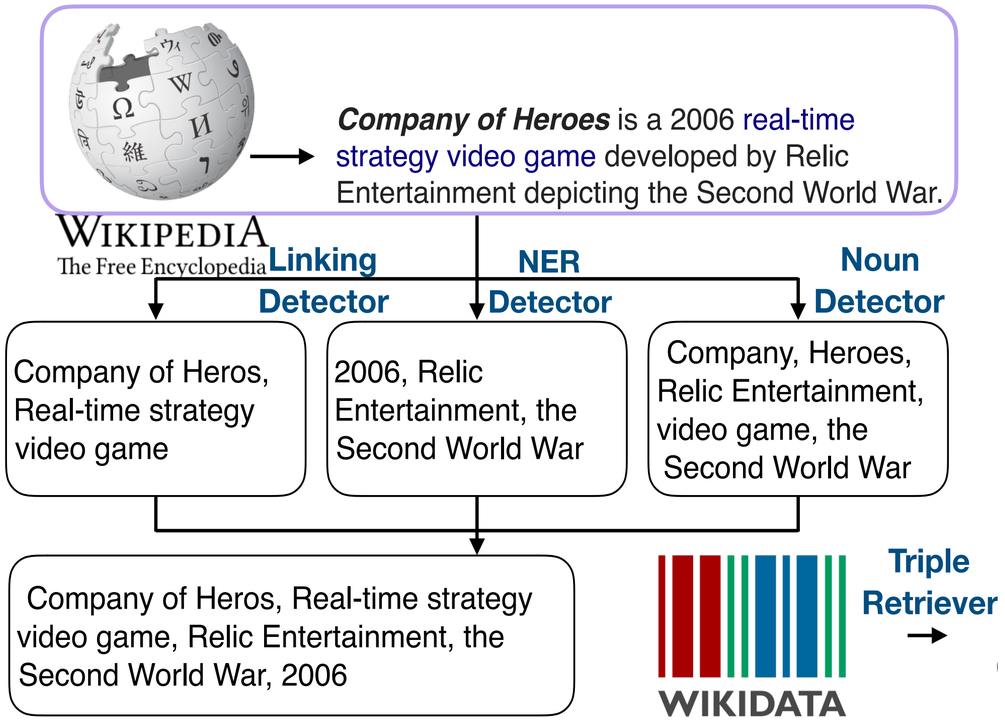
- Store the Wikidata database into Elasticsearch
- Search key:

  (head entity, tail entity)
- Return Value:

  <head entity, relation, tail entity>

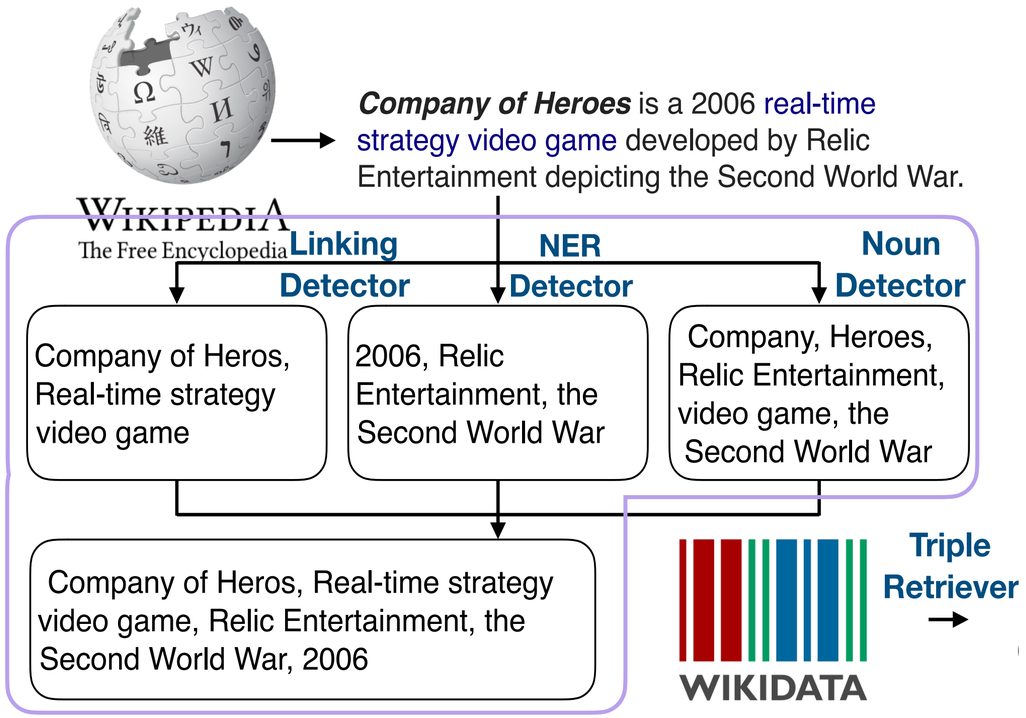
Key	Values
(Age of Empires, strategy video game)	<Age of Empires, genre, strategy video game>
(Bill Gates, Seattle)	<Bill Gates, birth_place, Seattle>
(Company of Heros)	<Company of Heros, publication_date, 2006>
…	…

Dataset - Sample sentences



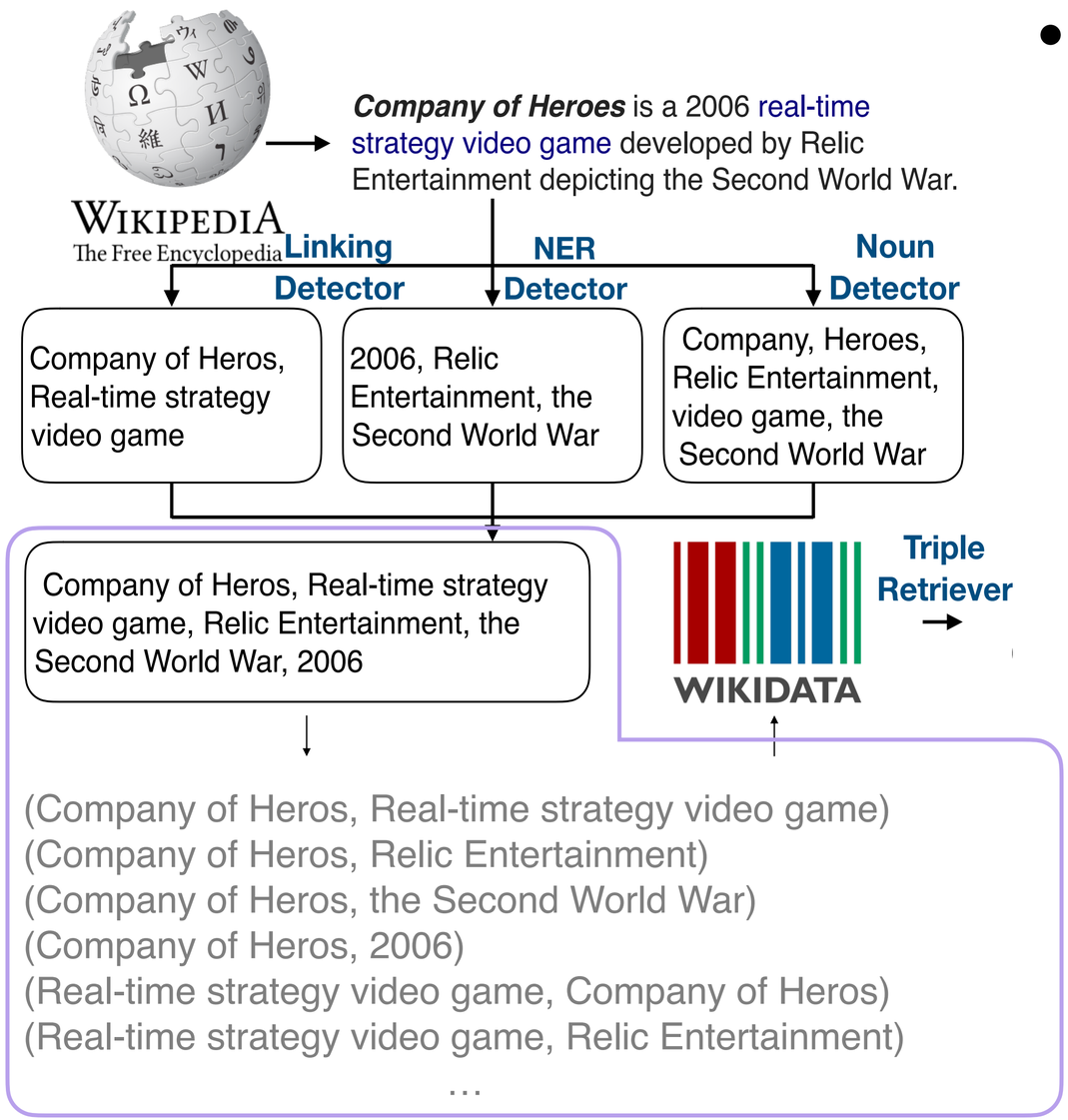
- We select each Wikipedia article's first sentence as the target text;
- Then, we remove irrelevant tags and links with several pre-defined rules.

Dataset - Entity Detector



- The entity detector contains three sub-detectors to recognize named entities and union them together as E_c.
  - Linking Detector: rule-based, can extract entities tagged with internal links;
  - NER Detector: Use spaCy's NER tool;
  - Noun Detector: based on spaCy's noun chunks recognition component to identify noun chunks.

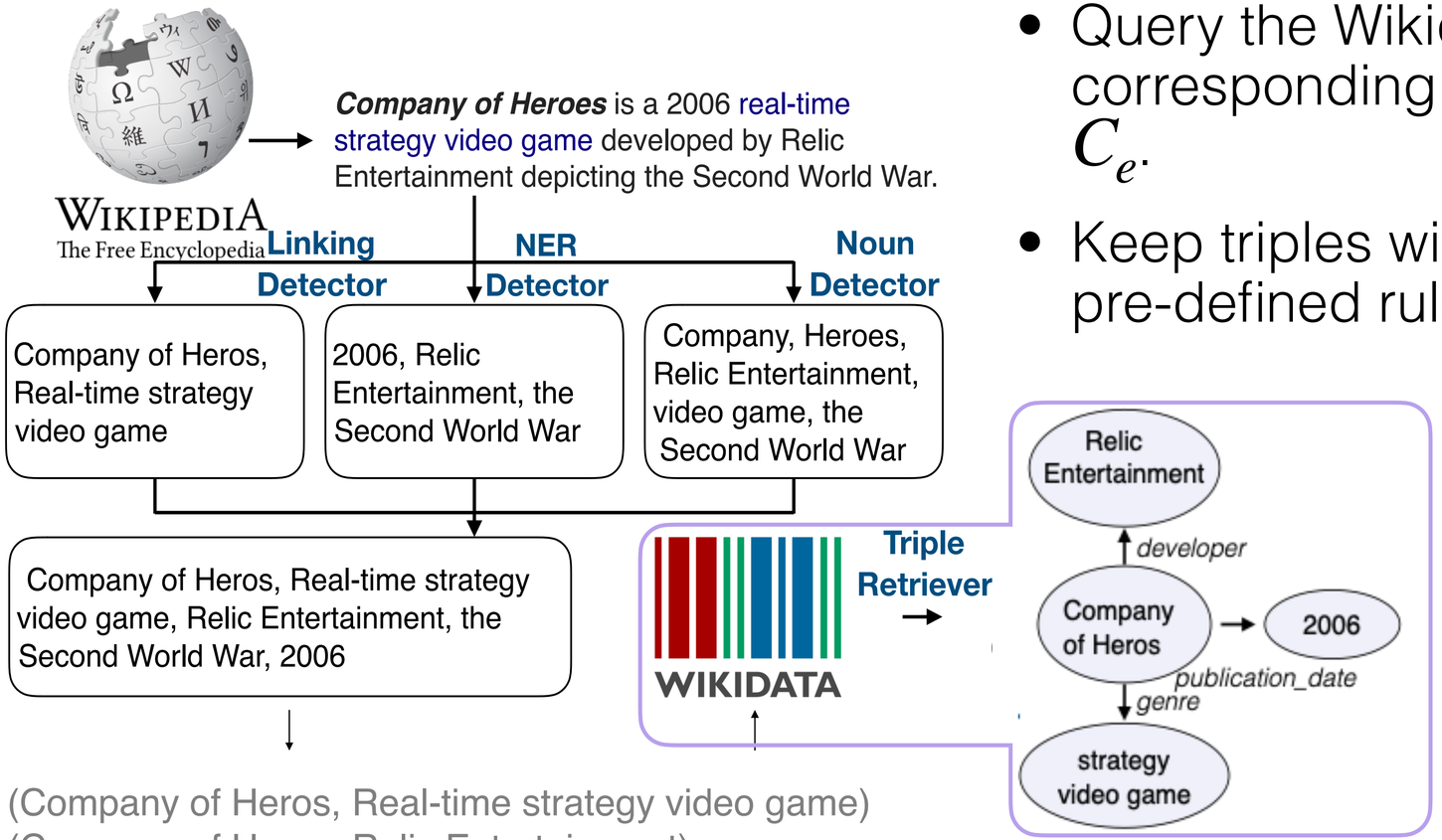
Dataset - Triple Retriever



- For given detected entities E_c, we make a list of named entity pairs by conducting a Cartesian Product as:

$$C_e = \{\langle e_i, e_j \rangle \,|\, \forall e_i \in E_c, e_j \in E_c, e_i \neq e_j\}$$

Dataset - Triple Retriever



(Company of Heros, Real-time strategy video game)
(Company of Heros, Relic Entertainment)
(Company of Heros, the Second World War)
(Company of Heros, 2006)
(Real-time strategy video game, Company of Heros)
(Real-time strategy video game, Relic Entertainment)
…

- Query the Wikidata database to find the corresponding triple for each entity pair in C_e.
- Keep triples with high scores with some pre-defined rules.

Dataset - Our WITA dataset

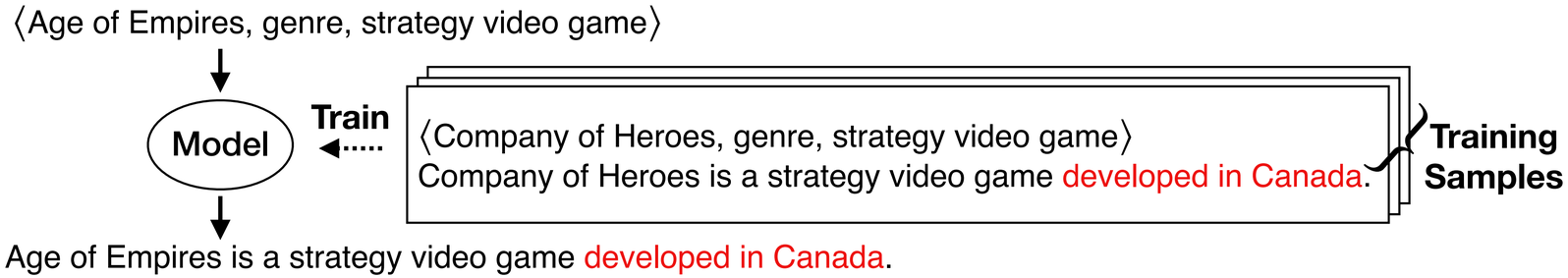
- (1) WITA is larger than the WebNLG dataset making it more practical.
- (2) WITA contains more relation types and entity types than that of WebNLG, indicating that our dataset involves more domains.
- (3) The vocabulary of the target sentences of WITA is much larger than that of WebNLG, which shows that our dataset is more challenging and more realistic.
- (4) It is automatically built and thus can be easily extended to other domains.

	WITA	WebNLG
Size	55,400	42,892
Relation Type	640	373
Entity Type	128,405	3,114
Text Length	(18.8, 17.0, 5, 59)	(23.7, 22.0, 4, 84)
KB Number	(3.0, 3.0, 1, 11)	(2.9, 3.0, 1, 7)
Vocabulary	102,404	8,886
entity-recall	0.508	0.625

Table 1: Statistics of WITA and WebNLG. For the text length and KB number, the data are mean, median, min and max respectively.

Dataset - Over-Generation Problem

- The automatically generated dataset is partially-aligned. Some text has no corresponding knowledge triple.
- Directly using the dataset to train the model makes the model bind unrelated text to the knowledge triple.
- When similar triples exist in the testing phase, it is prone to wrongly add some over-generated text, which is actually unrelated to the given input data.



Outline

✓Introduction

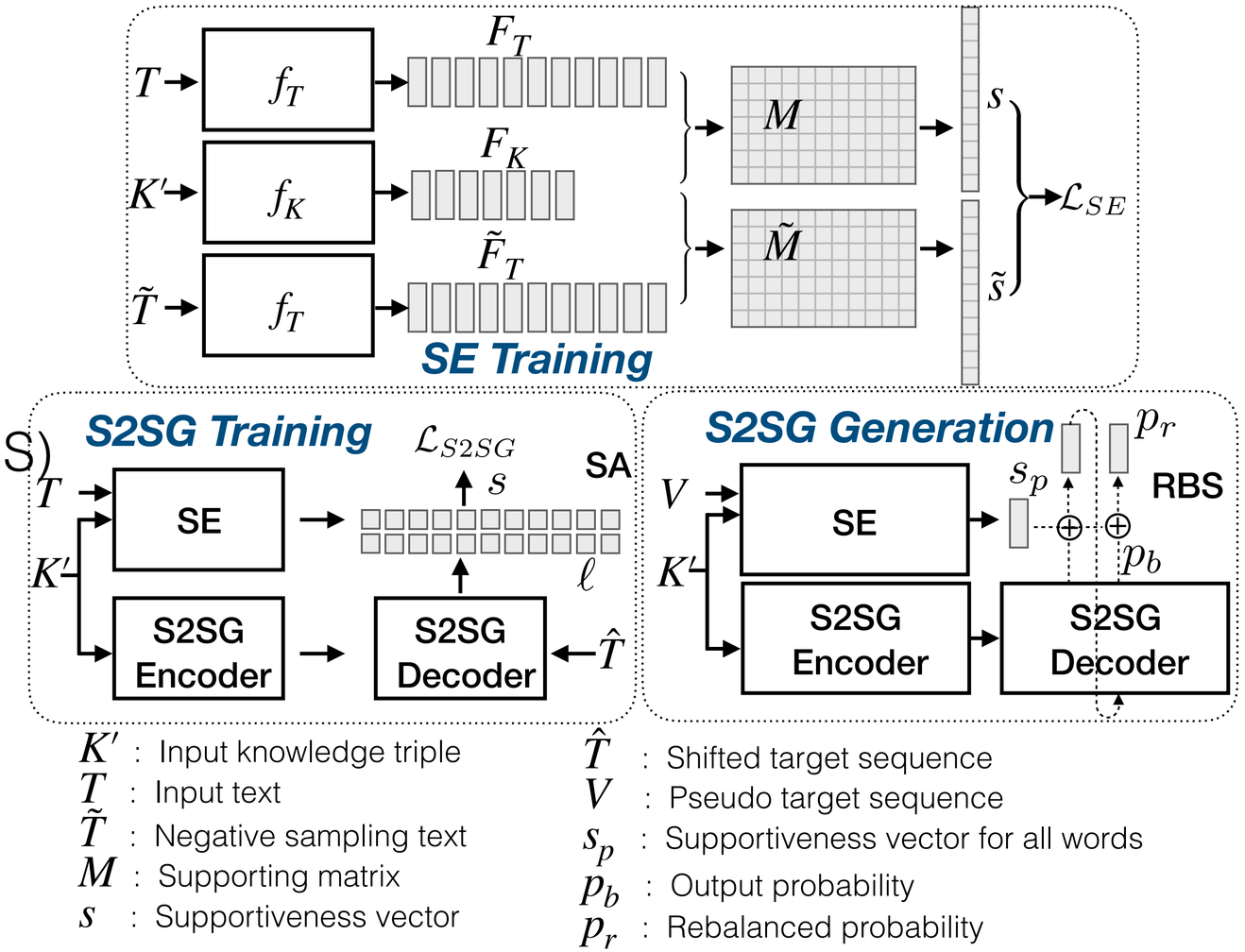
✓Dataset

➡Model

- Experiments
- Conclusions

Framework

- Our Distant Supervision Generation (DSG) framework contains four parts:
  - Supportiveness Estimator
  - Sequence-to-Sequence Generator (S2SG)
  - Supportiveness Adaptor (SA)
  - Rebalanced Beam Search (RBS)



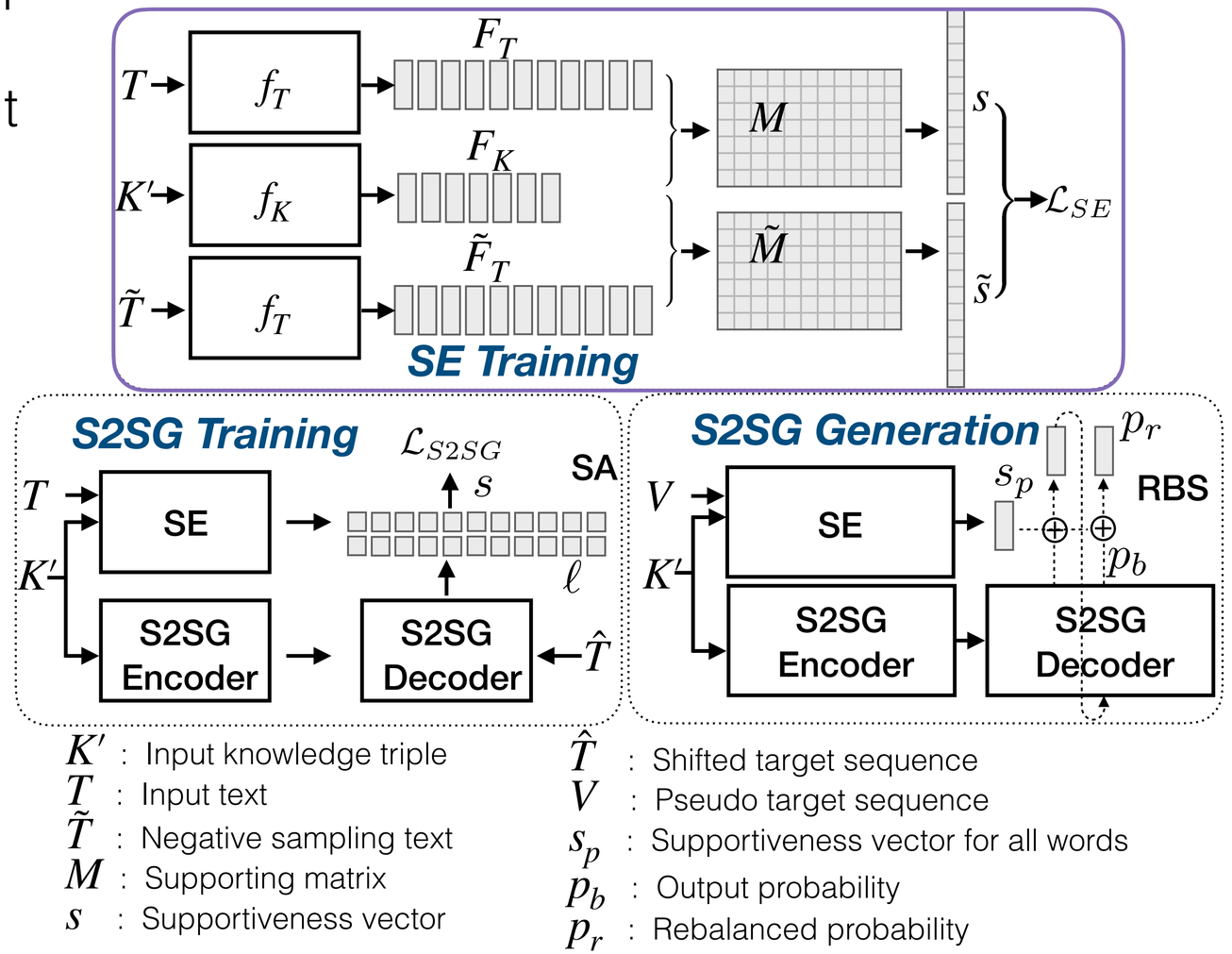
Framework - Supportiveness Estimator

- Supportiveness Estimator (SE) estimates a supportiveness vector $s \in \mathbb{R}^m$ indicating whether each target word w_i describes the input knowledge triples K'.
- Extract feature:

$$F_T = f_T(T);$$
$$\tilde{F}_T = f_T(\tilde{T});$$
$$F_K = f_K(K').$$

- Supporting matrix:

$$M = F_K^T F_T;$$
$$\tilde{M} = F_K^T \tilde{F}_T.$$



K' : Input knowledge triple
T : Input text
$\tilde{T}$: Negative sampling text
M : Supporting matrix
s : Supportiveness vector
$\hat{T}$: Shifted target sequence
V : Pseudo target sequence
s_p : Supportiveness vector for all words
p_b : Output probability
p_r : Rebalanced probability

Framework - Supportiveness Estimator

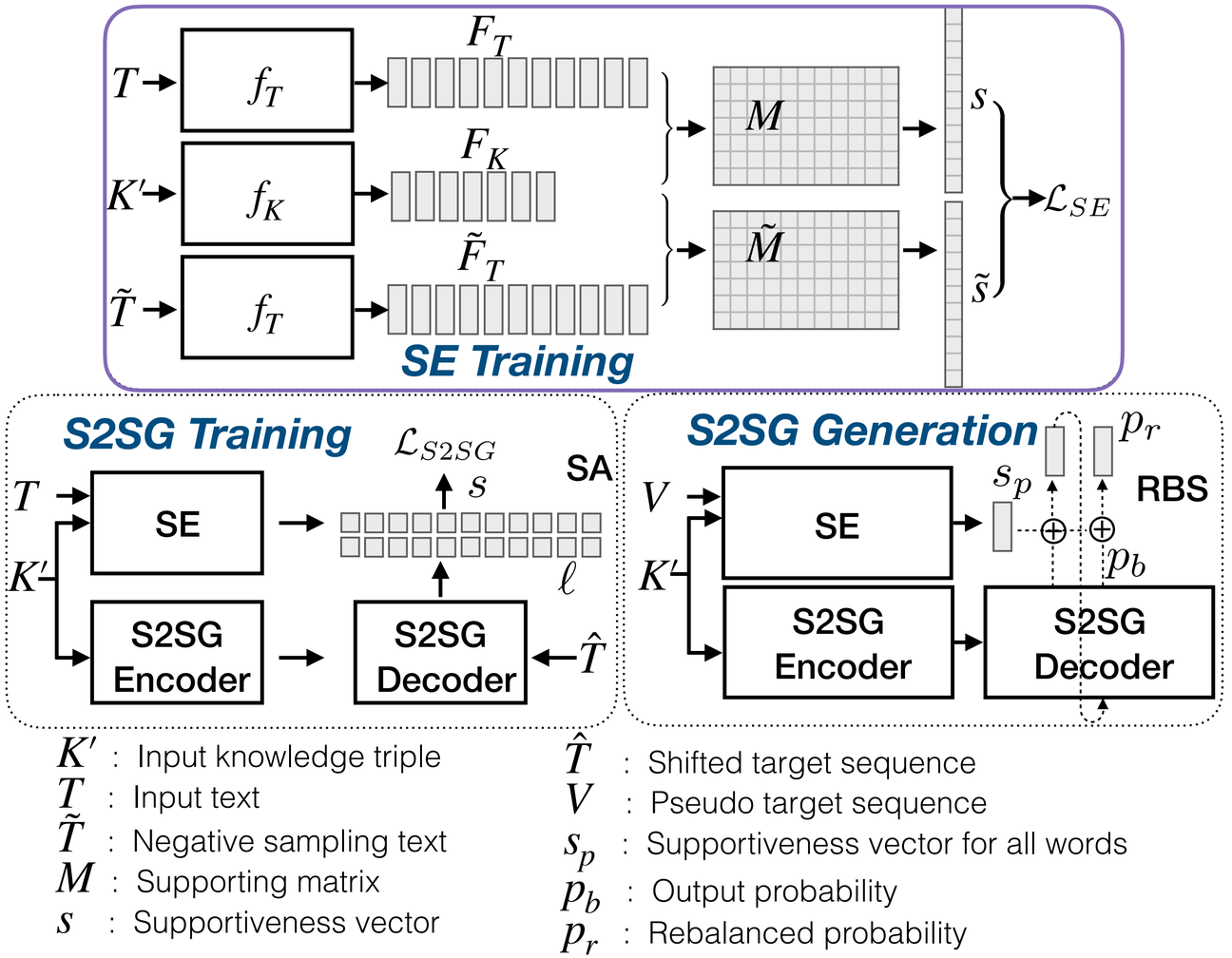
- Supportiveness score vector:

$$s_j = \log \sum_{i=1}^{|K'|} \exp(M_{i,j})$$

$$\tilde{s}_j = \log \sum_{i=1}^{|\tilde{K}'|} \exp(\tilde{M}_{i,j})$$

- Optimization Target:
  - margin loss:

$$\mathcal{L}_m = \sum_{i=1}^{\tilde{m}} \sigma(\tilde{s}_i) - \sum_{i=1}^{m} \sigma(s_i)$$



K' : Input knowledge triple
T : Input text
$\tilde{T}$: Negative sampling text
M : Supporting matrix
s : Supportiveness vector
$\hat{T}$: Shifted target sequence
V : Pseudo target sequence
s_p : Supportiveness vector for all words
p_b : Output probability
p_r : Rebalanced probability

Framework - Supportiveness Estimator

- word consistent loss:

$$\mathcal{L}_w = -\sum_{i=1}^{m}\sum_{j=1}^{|K'|} 1(T_i = K'_j)$$

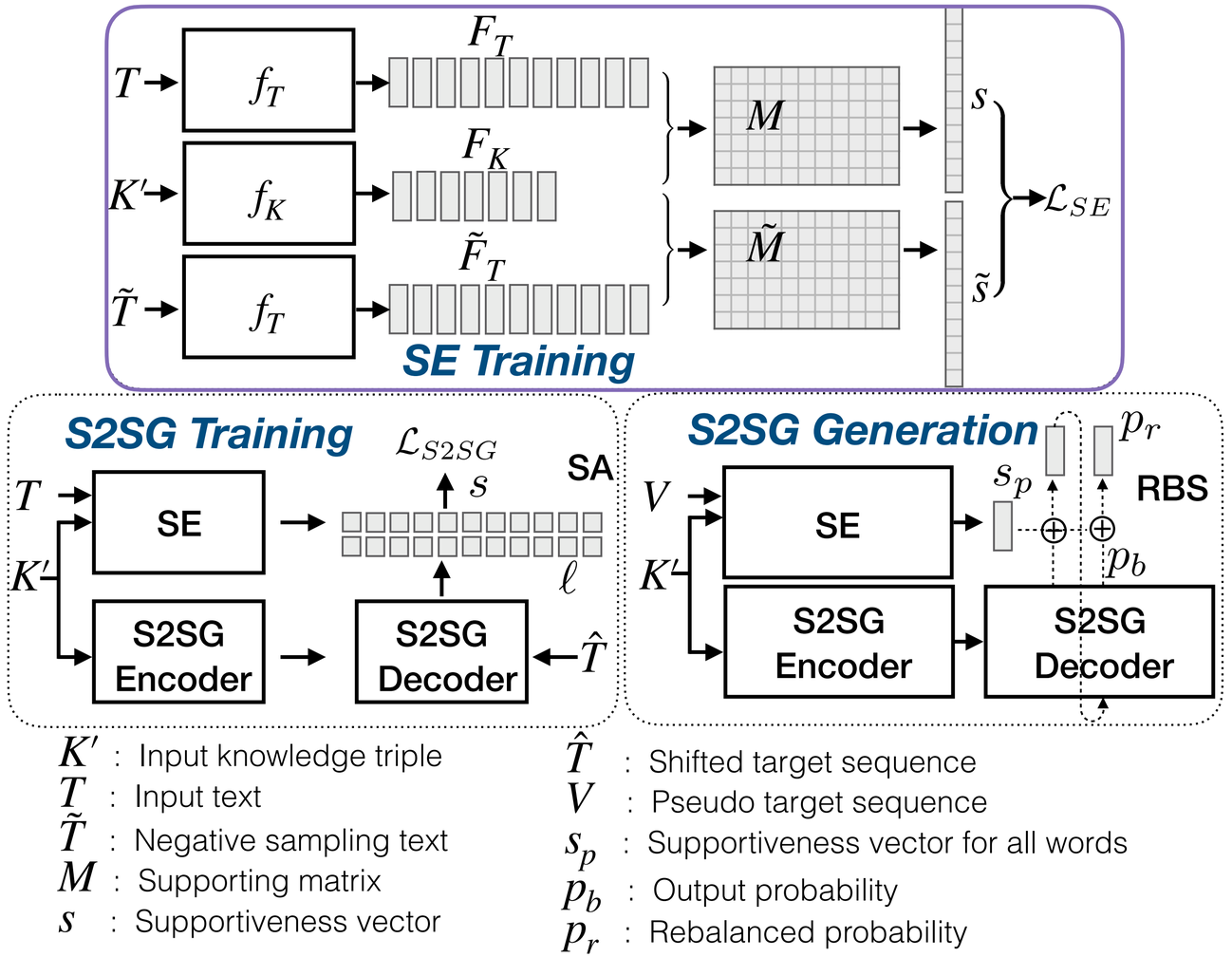
$$[M_{i,j} - \log(\sum_{k=1}^{|K'|} \exp M_{k,j})]$$

- concentration loss:

$$\mathcal{L}_c = \max_i \sum_{j=1}^{m} M_{i,j}$$

- overall loss:

$$\mathcal{L}_{SE} = \mathcal{L}_m + \omega_w \mathcal{L}_w + \omega_c \mathcal{L}_c$$

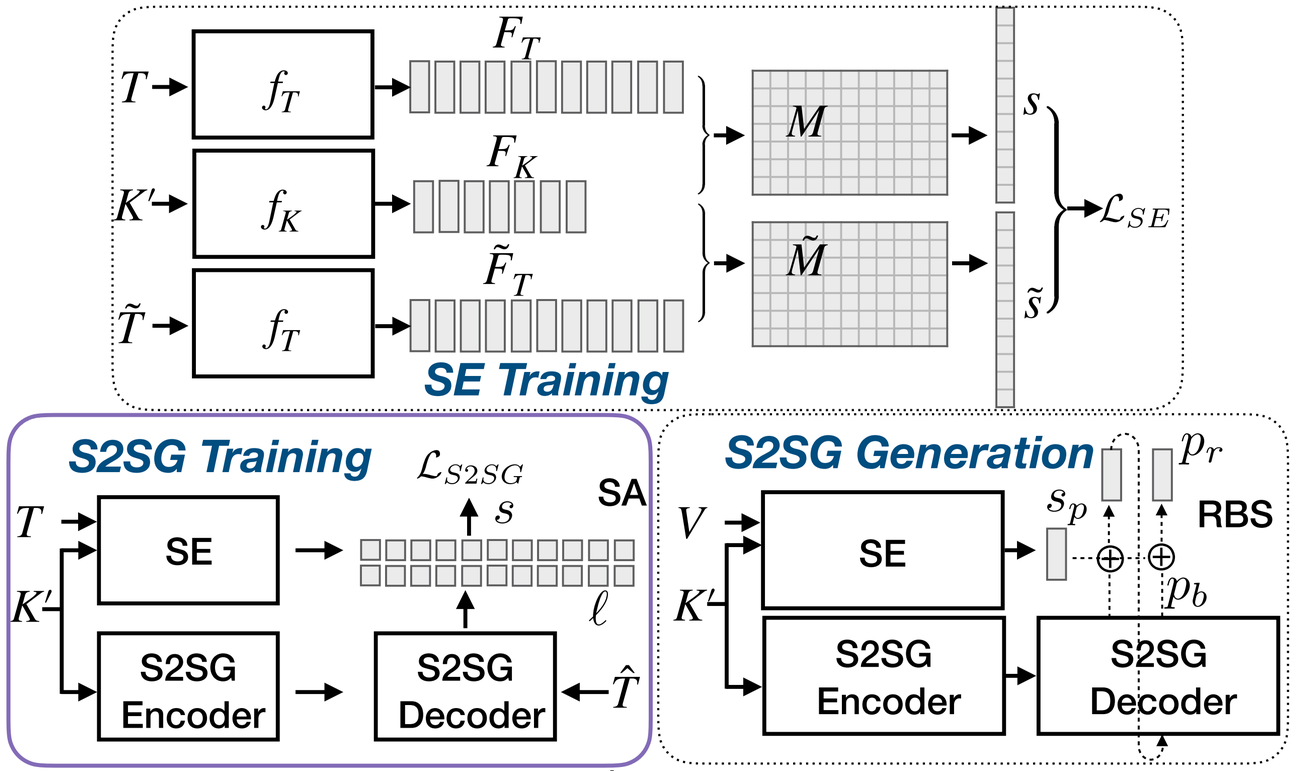


Framework - Sequence-to-Sequence Generator

- Sequence-to-Sequence Generator (S2SG) generates text based on the input sequence.
- It uses a Transformer:

$$G_K = \mathrm{Enc}(K')$$

$$\ell = \mathrm{Dec}(\hat{T}, G_K)$$



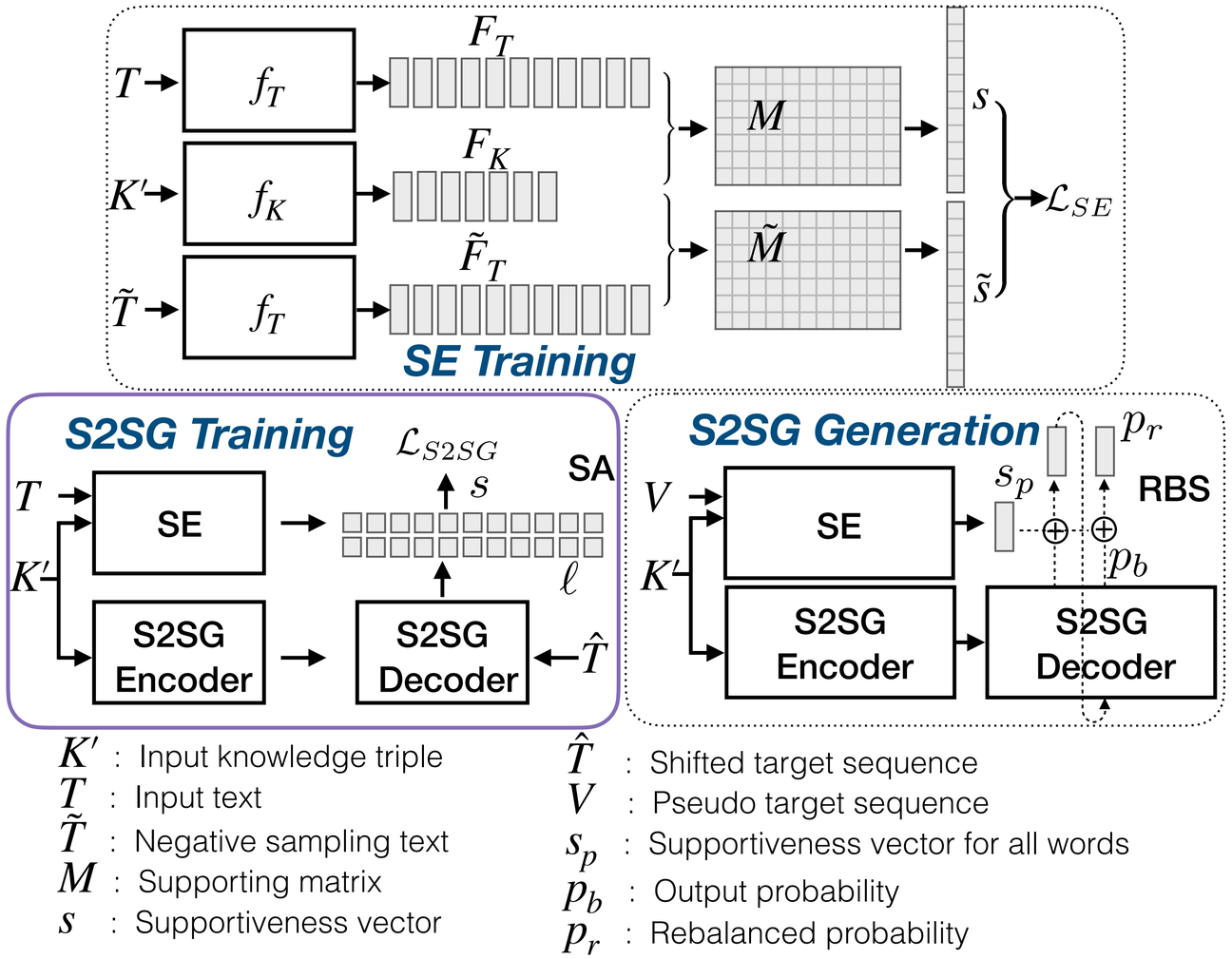
K' : Input knowledge triple
T : Input text
$\tilde{T}$: Negative sampling text
M : Supporting matrix
s : Supportiveness vector
$\hat{T}$: Shifted target sequence
V : Pseudo target sequence
s_p : Supportiveness vector for all words
p_b : Output probability
p_r : Rebalanced probability

Framework - Supportiveness Adaptor

- Supportiveness Adaptor adapts the supportiveness score s to S2SG's output ℓ.

$$\mathcal{L}_{S2SG} = \sum_{i=1}^{m} \ell_i s_i$$

- A target word's loss will be negligible if its supportiveness score is small.

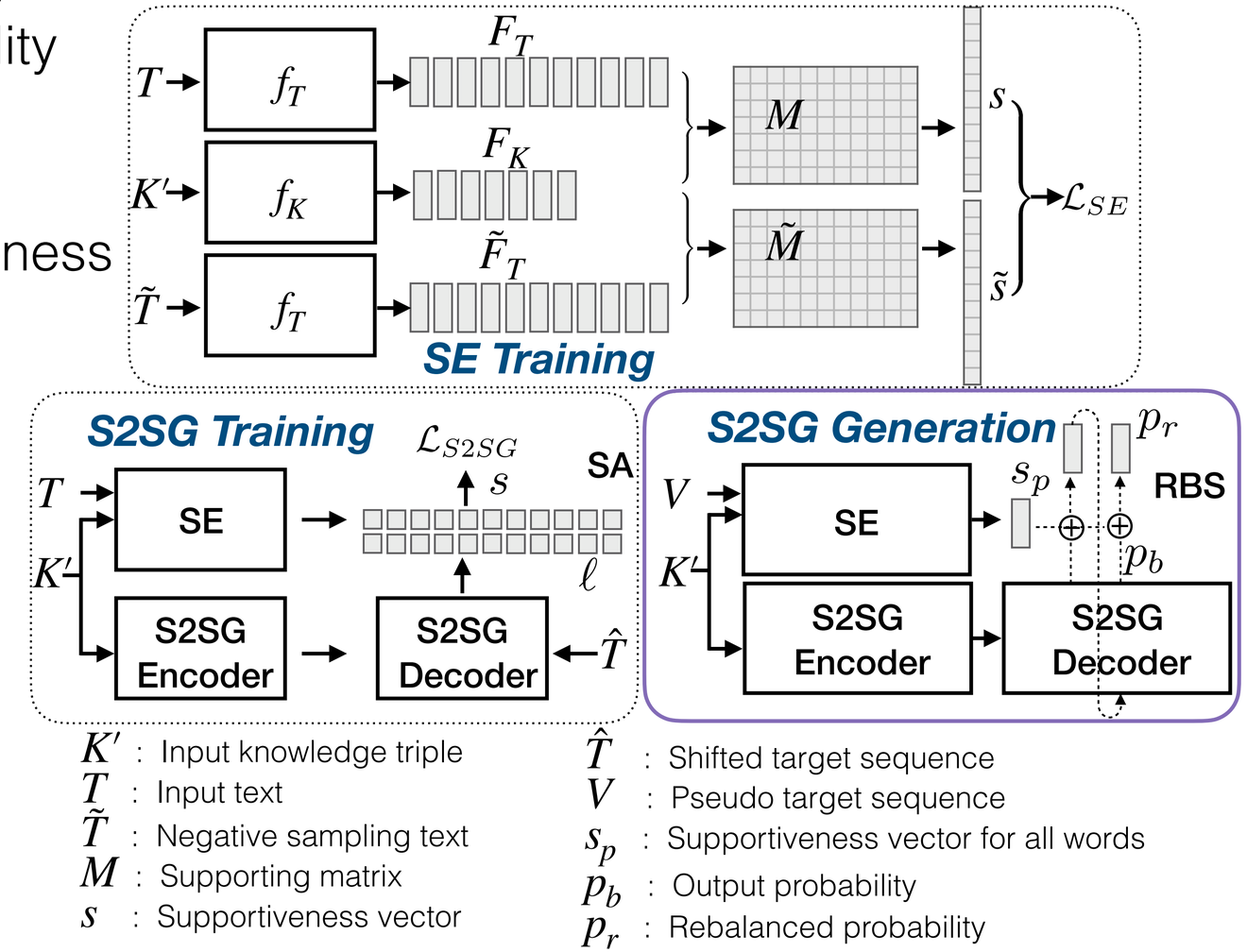


K' : Input knowledge triple
T : Input text
$\tilde{T}$: Negative sampling text
M : Supporting matrix
s : Supportiveness vector
$\hat{T}$: Shifted target sequence
V : Pseudo target sequence
s_p : Supportiveness vector for all words
p_b : Output probability
p_r : Rebalanced probability

Framework - Rebalanced Beam Search

- In the generation step, the supportiveness scores is used to rebalance the final word probability distribution.
- For the output probability $p_b \in \mathbb{R}^{|V|}$,we use the supportiveness score $s_p \in \mathbb{R}^{|V|}$ for all word to rebalance it:

$$p_r = p_b \cdot s_p^{\alpha}$$



Outline

✓Introduction

✓Dataset

✓Model

➡Experiments

• Conclusions

Experimental Results

- Our proposed DSG model performs the best to handle the partially-aligned dataset;
- Our proposed model alleviates the over-generation problem;
- Our proposed RBS & SA components both help improve the performance.

	BLEU	NIST	METEOR	ROUGE$_L$	CIDEr
S2S	0.463	7.97	0.385	0.693	4.12
S2ST	0.496	8.05	0.417	0.721	4.53
DSG-A	0.518	8.36	0.421	0.730	4.75
DSG-H	0.500	8.61	0.403	0.711	4.65
DSG	**0.555**	**8.71**	**0.425**	**0.742**	**5.02**
DSG w/o RBS	0.540	8.59	0.421	0.740	4.97
DSG w/o SA	0.522	8.38	0.421	0.734	4.83

Table 2: Main results.

	1-gram	2-gram	3-gram	4-gram	5-gram
S2ST	962	2,313	3,118	3,425	3,501
DSG-A	894	2,161	2,934	3,217	3,290
DSG-H	646	1,817	2,494	2,786	2,854
DSG	741	1,894	2,599	2,870	2,936

Table 3: N-gram statistics for over-generation error analysis.

Experimental Results

- Our new model generates more human-readable text;
- The supportiveness estimator correctly gives each target word a supportiveness estimation.

	Overall	Match
S2ST	7.315	7.231
DSG-H	7.285	7.331
DSG	7.377	7.569

Table 6: Human evaluation.

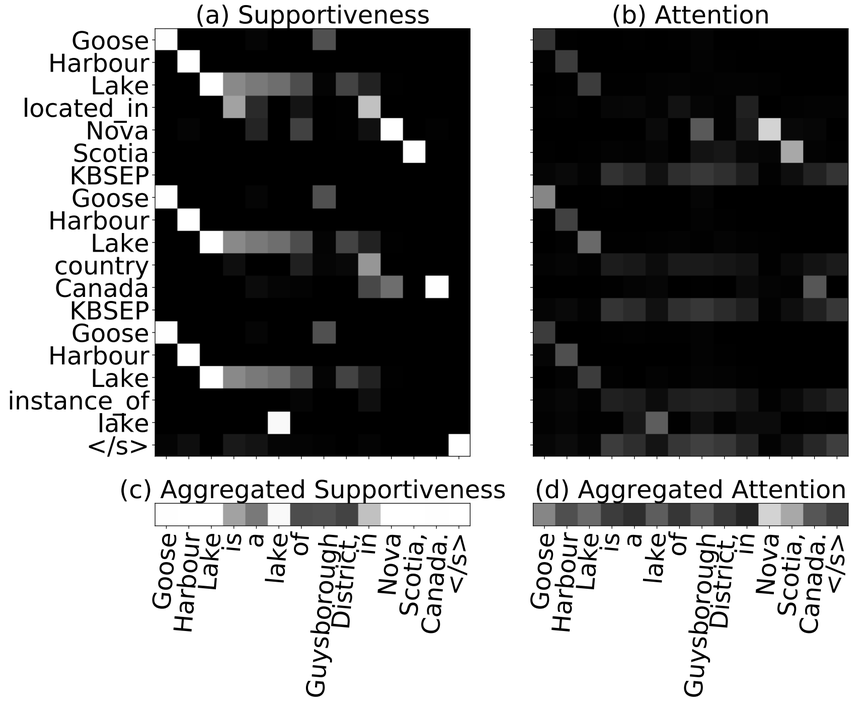


Figure 4: Comparision for supportiveness and attention. x-axis is the target text while y-axis is the given input KB triples. White stands for high score while black stands for low score. (a) and (b) show how each word in KB triples and text is aligned. (c) and (d) show the aggregated supportiveness and attention for each word in the target text.

Experimental Results

KB Triple	S2ST	DSG-H	DSG	Gold
⟨Four Crowned Martyrs, genre, sculptural group⟩, ⟨Nanni di Banco notable_work, Four Crowned Martyrs⟩	The Four Crowned Martyrs (also known as the Four Crowned Martyrs) is a sculptural group four by Nanni di Banco.	"Four Crowned Martyrs" a sculptural group Nanni di Banco.	Four Crowned Martyrs is a sculptural group by Nanni di Banco.	Four Crowned Martyrs is a sculptural group by Nanni di Banco.
⟨Newfoundland and Labrador Route 341, located_in, Newfoundland and Labrador⟩	Route 341 is a rural road in the Canadian province of Newfoundland and Labrador.	Route 341 a Canadian of Newfoundland and Labrador.	Route 341 is a settlement in Newfoundland and Labrador.	Route 341 is located in in Newfoundland and Labrador.
⟨Gaius Helen Mohiam, creator, Frank Herbert⟩, ⟨Gaius Helen Mohiam, instance_of, fictional character⟩, ⟨Dune universe, creator, Frank Herbert⟩, ⟨Dune universe, instance_of, fictional universe⟩, ⟨ Gaius Helen Mohiam, from_fictional_universe, Dune universe⟩	Gaius Helen Mohiam is a fictional character appearing in American comic books published by Frank Herbert.	Gaius Helen Mohiam is a fictional character created by Frank Herberfor the Dune univer.	Gaius Helen Mohiam is a fictional character in the Dune universe stationed by Frank Herbert.	Gaius Helen Mohiam is a fictional character in the Dune universe created by Frank Herbert.

Table 5: Case study. The red font stands for over-generated words while the blue underline indicates incoherent parts.

Outline

✓Introduction

✓Dataset

✓Model

✓Experiments

➡Conclusions

Conclusions

- We propose a new task, namely, partially-aligned Data-to-Text generation;
- We contribute a partially-aligned dataset WITA;
- We propose a distant supervision generation framework that successfully solves the task.

Thank You

https://github.com/fuzihaofzh/distant_supervision_nlg